Trends in Cognitive Sciences

CellPress

Opinion

A high-dimensional model of social impressions

Jonathan B. Freeman [1,*] and Chujun Lin [1]

People form social impressions from visual cues such as faces, which are argued by various models to arise from some limited set of fixed dimensions (e.g., trustworthiness and dominance). We argue that these dimensions, rather than reflecting intrinsic mechanisms, emerge from adaptive visuo-semantic processes in a high-dimensional neural-state space. Drawing on attractor neural-network models, we propose a framework treating social impressions as dynamic trajectories that stabilize over time, influenced not only by visual cues but also by conceptual associations and higher-order social cognition. Unlike low-dimensional models, this framework can account for cultural, individual, and situational factors that shape impressions. A high-dimensional framework makes several novel predictions and can offer a more accurate and complete understanding of the fluidity and complexity of social perception.

Highlights

Standard low-dimensional models summarize face impressions using a small set of fixed dimensions, such as trustworthiness and dominance. However, these reduced dimensions are increasingly found to vary across targets, perceivers, and methods, highlighting the need for a more dynamic framework.

We propose a high-dimensional model, where impressions arise from distributed neural patterns shaped by facial features, conceptual associations, and social cognitive processes. This approach resolves inconsistencies in low-dimensional findings, including cultural, individual, and situational variability, emphasizing that dimensions are emergent patterns rather than fixed mechanisms.

We do not believe a universal low-dimensional model exists. A high-dimensional framework offers a more accurate and complete understanding of face impressions and can provide deeper insights into the flexibility of social judgments.

Social impressions

People make any number of social inferences in just the blink of an eye, which occur rapidly and outside awareness [1,2]. These judgments often arise from 'facial stereotypes' and are largely inaccurate. Nevertheless, they shape decision-making in a wide range of contexts, including hiring, voting, courtroom sentencing, dating, and financial decisions [3,4].

A key insight over the past 15 years is that the various impressions people form from faces can be summarized by a small number of dimensions. Early work reported two dimensions – trustworthiness and dominance [5] – consistent with a larger body of work on the 'Big Two' dimensions in social perception more broadly, such as warmth and competence [6] and other models [7–9]. However, subsequent studies using different faces and traits found an increasing number of dimensions (Box 1). For instance, using ambient faces with a wider range of ages, researchers reported three dimensions: approachability, dominance, and attractiveness/youthfulness [10]. By further maximizing the diversity of stimuli and traits, other researchers found four dimensions: warmth, competence, femininity, and youthfulness [11]. Some studies have reported even 5–14 dimensions [12,13]. Common to all these models is the assertion that a small number of dimensions can capture substantial variation in impressions. Here we refer to these perspectives collectively as low-dimensional models.

The ever-increasing set of low-dimensional models has provided an elegant framework for researchers to study impressions. Given a limited set of dimensions discovered through dimensionality-reduction techniques like principal component analysis (PCA) or factor analysis, researchers can focus exclusively on measuring judgments of the model's dimension labels (e.g., trustworthiness) to answer a range of questions. Since the underlying dimensions are assumed to summarize most variance in people's judgments, researchers can focus exclusively on measuring the latent dimensions without needing to concern themselves with countless

[1]Columbia University, New York, NY, USA

*Correspondence: jon.freeman@columbia.edu (J.B. Freeman).

Box 1. Low-dimensional models of face impressions

The number of low-dimensional models is growing rapidly. These models often begin with unconstrained free responses or dictionary lists, followed by validation of the dimensions with constrained judgments. The large variation of low-dimensional models presents challenges to the field in obtaining a unified understanding of the mechanisms underlying face judgments. However, common to all these models is the premise that there is a set of relatively fixed core dimensions that summarize the majority of variance in face impressions (Figure I). Moreover, certain dimensions (e.g., trustworthiness/approachability/warmth) can often reappear across different models, albeit using different labels. Variations in these models are mainly due to three sources: targets, perceivers, and experimental methods.

Targets: First, by varying the targets whose face images are shown, different dimensions are found for judgments of children's faces (two dimensions of niceness and shyness) [46], older adult faces (two dimensions of sternness and confidence) [47], and racially diverse faces (five dimensions of warmth, competence, femininity, youth, and race) [48]. Some studies have also added more naturalistic information, such as using dynamically moving faces in videos, which reveal seven and even 25 dimensions [16,49].

Perceivers: Second, by varying the perceivers who make the judgments of the faces, prior work found different dimensions for participants from Western cultures (approachability and dominance) than participants from East Asian cultures (approachability and capacity – the latter of which is more related to judgments of intelligence and attractiveness) [50]. Different numbers of dimensions are also found for perceivers from different world regions, ranging from two to four) [15].

Methods: Varying the methods for collecting face judgments and conducting dimensionality-reduction analysis also reveals different dimensions. For instance, while past work reporting four dimensions assessed ratings of faces on 100 maximally representative English trait words, more recent work that asked participants to freely arrange the faces spatially based on their perceived personality (without prompting participants on any specific preselected traits) revealed five dimensions [13]. The five dimensions likely reflect a more holistic evaluation of faces when judgments are not artificially constrained by specific traits. Similarly, recent studies using free-response data with natural language processing and factor analysis found that 14 dimensions best explain judgments of Facebook profile photos, which include personality (e.g., sociability, adventurousness) and sociodemographic characteristics (e.g., gender, age, weight) [12]. Another study applied a different approach (clustering) to the dataset previously reporting four dimensions that used factor analysis, PCA, and artificial neural networks, now only finding two dimensions, approach and avoidance, rather than four [51].

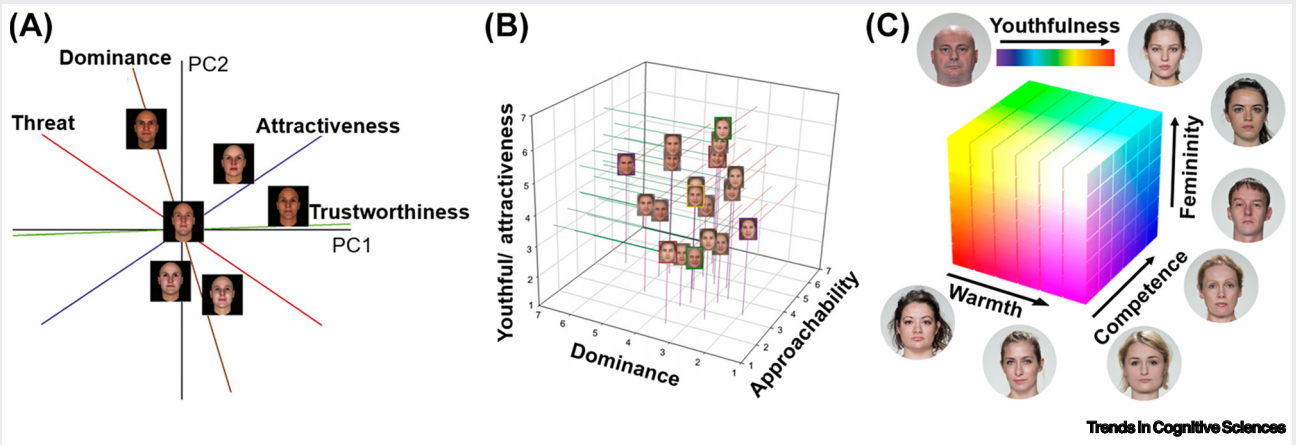

Figure I. The 2D, 3D, and 4D models of face impressions. (A) The 2D model of trustworthiness and dominance [5]. Abbreviation: PC, principal component. Note that while PC1 and PC2 are orthogonalized, trustworthiness and dominance traits are somewhat negatively correlated. Adapted from [5] with permission. (B) The 3D model of approachability, dominance, and youthfulness/attractiveness [10,52]. Adapted from [52] with permission. (C) The 4D model of warmth, competence, youthfulness, and femininity [11]. Face images are from the Face Research Lab London Set [53].

other traits [14]. These models generally focus on content rather than process, providing static statistical descriptions of social judgments once judgments are already made. However, core dimensions have often been interpreted in the literature as reflecting underlying cognitive or functional mechanisms. For instance, when the 2D trustworthiness–dominance model was first introduced over 15 years ago, the dimensions were described as having a functional basis in that they reflect evolutionarily adaptive cognitive mechanisms [5]: while trustworthiness reveals people's good or bad intentions, dominance reveals people's ability to enact those intentions. Some researchers have argued that intuiting such information may be a universal property of social cognition [6].

While not all proponents of low-dimensional models take this view [10,11], once such models are described, the literature tends to interpret the dimensions as reflecting distinct psychological

mechanisms. This can also be a theoretical assumption of the analyses used. Increasingly, researchers have favored factor analysis with oblique rotation over PCA, to allow dimensions to be correlated rather than force orthogonality [15]. This is sensible as face impressions' reduced dimensions naturally tend to be correlated (see Figure IA in Box 1). However, while PCA is theoretically agnostic in merely providing a convenient statistical solution for reduced dimensions, factor analysis makes a strong assumption that the latent constructs identified reflect underlying mechanisms that cause variation in judgments [15].

Although rarely articulated explicitly, at a process level, low-dimensional perspectives (particularly when supported by factor analyses) would view the architecture of social perception as having dedicated mechanisms for perceiving certain core dimensions (e.g., trustworthiness and dominance), which are then combined to form judgments of all possible traits [16]. As such, trait dimensions would be viewed as functioning somewhat similarly to color dimensions, where visual input is first encoded into red, green, and blue dimensions, after which different mixtures of these dimensions give rise to a full subjective color space. However, while in color perception we have long known that there are three types of cones in the retina that are selectively responsive to the three primary colors, the notion that a similar set of detector mechanisms are dedicated to core trait dimensions has not been tested.

Given increasing studies incorporating more diverse targets, perceivers, and conditions that show the dimensionality-reduction solutions for face impressions are highly variable and context dependent, we do not believe any universal low-dimensional model is possible. Nevertheless, we recognize that certain trait dimensions, such as trustworthiness/approachability/warmth, can emerge consistently across different datasets and contexts. Rather than conceptualizing these dimensions as fixed causes of impressions, we argue that they are manifestations of natural clusters of associated trait impressions that tend to consistently vary due to a number of factors, such as overgeneralization and statistical learning [4], social-conceptual knowledge [17,18], or chronic goals and other internal states [5,6]. Here, we argue that the commonly observed low-dimensional statistical structures for social impressions may simply be an emergent property of a high-dimensional neural-state space.

Social impressions as trajectories through high-dimensional space

Drawing on insights from attractor neural-network models [19], we propose a high-dimensional framework for social impressions. Attractor models are recurrent systems where patterns of neural activity stabilize into steady states (i.e., attractors). These models are widely used to simulate perception, memory, and decision-making by showing how neuronal populations converge on specific activity patterns corresponding to learned representations or cognitive states. Attractors represent stable outcomes, such as a perceptual category, memory recall, or decision state. In this vein, we can conceive of social impressions as encoded by distributed neural patterns in an attractor neural network, where the network is attracted to specific neural population states that correspond to specific combinations of trait judgments. Our model aims to understand how combinations of impressions emerge dynamically in real time. In this model, a stable attractor state does not correspond to a single trait judgment, but rather the combination of all possible trait judgments at once (e.g., high trustworthy, high affectionate, low intelligent). The model builds on the dynamic interactive (DI) theory of social perception but here focusing on trait impressions rather than social categories and emphasizing attractor dynamics of distributed neural patterns [20,21].

Attractor models are often tested using recurrent neural networks, as with prior DI models [20,21]. In such a network, the process of forming trait judgments reflects a dynamic evolution

of trait nodes' activation. If we consider cases where faces are judged on 100 traits (as in prior work [11]), the network would include 100 trait nodes. As shown in Figure 1A (Key figure), in response to a face, detected features (left) begin activating combinations of trait nodes (middle), whose activation is also modulated by top-down processes (right). Some facial features may have strong connections to trait nodes and activate them quickly. Trait-node activations continue to evolve as the early-activated trait nodes recurrently pass activations onto other conceptually associated trait nodes, where traits suggestive of one another will activate each other (and those in conflict will inhibit each other). Traits can also directly compete with one another via mutual inhibition, such as trustworthy and untrustworthy nodes, to give rise to continuous judgments (i.e., classic bipolar trait dimensions). Simultaneously, trait nodes' activation is modulated by internal top-down states. Over hundreds of milliseconds, the node activations stabilize on an attractor state that maximally satisfies the constraints in the visual input, top-down input, and internal network structure (relationships among all features, traits, and top-down processes). Trait judgments are then determined by the final activation of all trait nodes once the system stabilizes, reflecting a compromise of all inputs and constraints.

We refer to our model as high-dimensional in two respects. The first is that the network encompasses a large number of trait nodes, which form the dimensions in the model (rather than reduced dimensions from PCA or factor analysis). This is similar to other recent network-based approaches to trait representation in perceiving the self [22,23]. In Figure 1A, ten example trait nodes are visualized (blue color indicates stronger trait-concept activation, 0–100%). However, while trait nodes are valuable for modeling purposes, in reality distinct trait nodes do not exist in the brain. Instead, this recurrent-network representation is a simplification of the actual neural-state space in which the specific combination of all trait nodes' activation is encoded by a single high-dimensional neural pattern (e.g., 10 000 neurons). We therefore also refer to the model as high-dimensional in that it operates in high-dimensional neural-state space.

In Figure 1A, a small subset of this 10 000-dimensional pattern is shown as a heatmap of 625-dimensional neural activation. When presented with a face, the system's state (i.e., the 625-dimensional activation pattern) would seek to settle into a specific stable pattern based on the input and network's internal structure, reflecting the unique activation of the ten traits (an attractor state). How strongly the perceiver rates the face on any given trait is therefore determined by the high-dimensional neural pattern. In Figure 1B, heatmaps show an even smaller subset of this space, with 64-dimensional neural activation patterns associated with states in which a target face is maximally rated as highly trustworthy, caring, sociable, or responsible, respectively. For simplicity, a maximally active trait label is shown for each heatmap, but note that each neural pattern is associated with a specific combination of activation for all possible trait nodes. In classic models of face impressions, trustworthiness is a core dimension (e.g., via PCA) with caring, sociability, and responsibility strongly loading onto it [5]. The pattern for trustworthiness (top panel) (i.e., the PCA dimension from classic models) may operate in this framework as the broad overlap (i.e., an average/centroid) of its related, more granular traits in the three lower panels.

Figure 1C visualizes a 2D projection of high-dimensional neural-state space (e.g., 64, 625, or 10 000 neural activations), where each point reflects a unique neural state of trait activation across all possible traits. One can imagine a perceiver's combinatorial trait judgment as a trajectory through this 2D space (blue arrow), where the neural pattern settles into one of the attractors (i.e., final stable pattern). Impression ratings would map onto these trait-activation levels once

Key figure

Illustration of a high-dimensional model

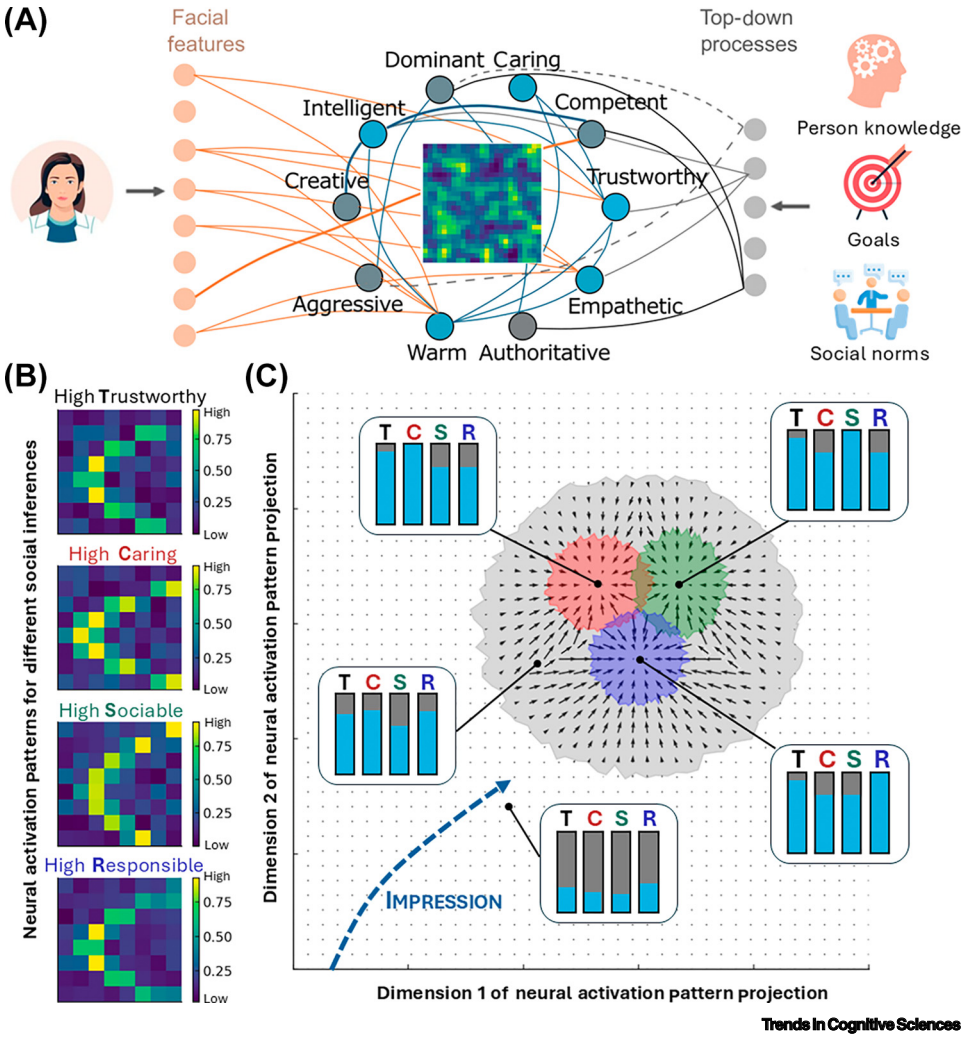

Figure 1. (A) A schematic of the network's architecture: facial features (left) activate related trait concepts (middle), which interact dynamically with top-down processes (right). Trait activations evolve through recurrent processing until the network settles into a stable configuration. While discrete trait nodes exist for purposes of network modeling, actual trait judgments emerge from distributed neural patterns (e.g., 10 000-dimensional activations), a simplified subset of which is shown here as a 625-dimensional heatmap. (B) Heatmaps of neural patterns in an even further simplified subset of 64-dimensional patterns are visualized, corresponding to states in which target faces are judged as maximally trustworthy (T), caring (C), sociable (S), or responsible (R). Although each panel highlights a maximally activated trait, each neural state reflects a specific combination of all trait activations. In this framework, the neural pattern for the classic trustworthiness dimension derived from principal component analysis (PCA) or factor analysis (top panel) reflects the average/centroid of related, more granular trait combinations (lower panels). (C) A 2D projection of the high-dimensional neural-state space. Each point reflects a unique combination of all trait activations (0–100%, blue color). Red, green, and blue attractor basins are visualized, corresponding to states where faces are maximally judged to be caring, sociable, or responsible, respectively. The gray attractor basin indicates how the trustworthiness dimension from PCA or factor analysis may function in this framework, as a kind of superordinate attractor that pulls the system into more granular representations.

the system reaches its final attractor state. Red, green, and blue attractor basins are visualized, whose central points reflect neural states where faces are maximally judged to be caring, sociable, or responsible, respectively.

In our model, putatively core dimensions (e.g., trustworthiness) do not have an intrinsically privileged status but merely reflect an emergent property of correlated neural patterns. Consider, for example, a face perceived to be caring, responsible, and sociable, traits strongly related to the classic trustworthiness dimension, as just discussed [5] (Figure 1B). After exposure to the face, the system's trajectory would begin converging on a distributed neural pattern associated with the exhaustive set of inferred traits (e.g., caring, responsible, sociable). As shown in Figure 1C, during this pattern-completion process, the system's state would increasingly sharpen into a pattern reflecting the unique activation of all traits [i.e., an attractor state (central points inside the green, blue, or red attractor basins)].

Because the correlation structure of these neural patterns is driven by the relatedness among trait concepts and visual features (due to the feature–trait and trait–trait network connections), correlated trait judgments will show more overlapping neural patterns. For instance, the caring, sociable, and responsible trait nodes are strongly positively connected, which will lead the partial activation of one of these nodes to trigger partial activation of the other two nodes. This mutually reinforced coactivation acts like a magnet, initially pulling the system toward a broad overlapping neural pattern (i.e., an average/centroid pattern) that resembles the 'trustworthiness' dimension from classic models (Figure 1B). This overlap creates a superordinate attractor basin, visualized in gray in Figure 1C, corresponding to a high-trustworthy evaluation. However, this superordinate attractor basin is not 'trustworthiness' *per se* but rather a temporary metastable state that reflects the overlapping partial activation across many correlated traits, before the system settles into a more granular attractor reflecting one's final impressions. Thus, while 'trustworthiness' may be identified by PCA or factor analysis, in the model it would reflect an emergent property of the correlation structure among traits rather than a fixed latent factor.

Social learning and top-down social cognitive processes

PCA and factor analysis have proved valuable for the identification of statistical structure in datasets. However, we argue that reducing dimensions can obscure important and systematic variations in face-impressions structure, which depend on who is judging, who is being judged, and a variety of contextual and social cognitive factors that situate impressions. For instance, much of the high-dimensional structure of impressions differs reliably across individual perceivers and cultures in terms of the pairwise relationships among traits (e.g., the extent to which aggressive-looking faces are judged intelligent and vice versa), even if data from most perceivers and cultures can converge on similar low-dimensional solutions [17,24].

While proponents of low-dimensional models have acknowledged the role of context (and sometimes tested its role), such models often cannot accommodate contextual effects without proposing a different low-dimensional structure (e.g., in the context of female faces, trustworthiness and dominance collapse into just a single dimension [25,26]). Dimensionality-reduction techniques also cannot specifically predict what dimensions will arise due to changes in facial features or conceptual and top-down processes. In a high-dimensional framework, however, these changes directly modify the attractor landscape and thus the correlated neural patterns giving rise to trait judgments. For example, attractor basins can become more overlapping or distinct based on the association of different trait-related neural states (Figure 2). If goals, expectations, or

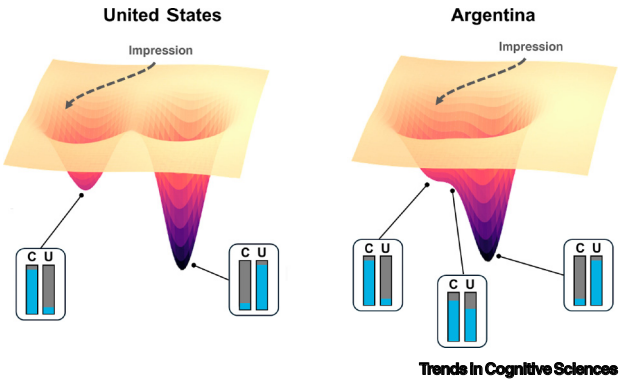

Figure 2. Dynamic modulation of attractor landscapes for social impressions. An illustration of how social cognitive processes can reshape attractor landscapes for face impressions. Landscapes indicate perceivers in the USA (left) and Argentina (right). The energy of the system is plotted as a function of a 2D projected view of high-dimensional neural-state space, where points correspond to specific combinations of all possible trait ratings in response to a face. Two attractors are visualized by local energy minima (i.e., stable states). One can imagine the current state of the system as a ball rolling down a hill, compelled to descend into one of the two local attractor states. Consider an example where, due to cultural learning, American perceivers tend to judge faces as either unhappy (U) or caring (C), with the traits relatively negatively correlated leading to distinct neural patterns (shown left). In Argentinian perceivers, however, the two traits are relatively more positively correlated due to cultural learning [29], leading associated attractor basins to blend and create a low-energy ridge where the system is able to stabilize in locations between the two attractors (and which therefore becomes attractive itself), creating more fluid dynamics involving stable states where faces can be judged as being both unhappy and caring (shown right). Note that the system's trait-activation states are associated with specific combinations of all possible traits, not just (U) and (C). This exemplifies how the attractors may change in topography to facilitate certain perceived traits or restructure those traits – whether due to attention, conceptual beliefs, social expectations, cultural learning, goals, stereotypes, or attitudes.

stereotypes heighten the salience of particular traits, the neural states emphasizing those traits may become more dominant in the network's dynamics, deepening the attractor basins most strongly related to particular trait combinations. A high-dimensional framework can therefore predict what specific low-dimensional structure will emerge based on particular cues, perceivers, and situations.

Perceivers' conceptual associations play a central role in our model. Studies have estimated a sizable role of perceiver characteristics in face impressions, including perceivers' conceptual understanding of different traits [24,27,28]. Individual differences in the extent to which any given pair of traits are deemed more conceptually similar (e.g., agreeable and open-minded) predicts a corresponding visual similarity in the facial features used to infer those traits [17,24]. Using data from over 40 countries, PCA or factor analysis found two or three dimensions that were generally consistent but also showed some cross-cultural variability, which was attributed to measurement error [15]. However, closer examination using high-dimensional analyses found that the structure of impressions differed systematically depending on a culturally learned conceptual structure of traits. For instance, while Americans tend to judge unhappy-looking faces as aggressive, Argentinians tend to judge such faces more as caring, and this could not be explained by mere linguistic differences. Instead, such cross-cultural differences in inferring traits from faces were related to differences in how those traits covaried in the actual personalities of people living in the local region, which perceivers likely encoded in the form of conceptual associations [29]. As the attractor landscape for impressions is built on both featural and conceptual relatedness among traits, such differences in learned conceptual associations about how traits are related intrinsically shape the landscape (Figure 2).

These conceptual associations also help to explain the theoretically infinite set of social attributes on which we are able to judge others based on facial appearance – even when the attributes lack any direct featural basis. Different traits show varying levels of interrater agreement, with some showing strong agreement (e.g., youthful) while others show weak agreement (e.g., creative). High-agreement traits are likely to be more 'proximal' to visual input and have stronger, more

direct associations with facial features. Such traits may thereafter activate more 'distal' traits that are only indirectly associated with visual input through trait conceptual associations [27,30]. Numerous factors are thought to drive the mappings among features and traits, including emotion overgeneralization, functional utility, individual and cultural learning [31], and network centrality [22,23]. Our model additionally suggests that, regardless of their origin, traits with more direct featural mappings can then propagate activation onto increasingly abstract social inferences that are less tethered to visual cues. For instance, as shown in Figure 1A, judgments of how creative a person is based on their facial appearance may not be directly associated with any specific features; instead, more proximal traits with a stronger featural basis (e.g., competent) may cascade activation via indirect conceptual associations (competent → intelligent → creative). Indeed, recent work has shown that such an indirect conceptual cascade helps to explain highly ambiguous social inferences based on facial appearance, such as education level and sexual orientation among others [30].

Top-down processes like goals, expectations, or social norms can shape impressions. For instance, trustworthiness is more closely tied to facial typicality for own-culture faces while more closely tied to facial attractiveness for other-culture faces [32]. Close and trusted individuals tend to elicit more positive evaluations of traits like dominance, which may be seen as a sign of strength, while dominance in distant or outgroup members may be registered as threatening. Accordingly, dominance and trustworthiness are more positively correlated when evaluating close others but more negatively correlated when judging unfamiliar or distant others [33,34].

Social categories and intergroup impressions

While face-impressions research historically focused on young White male faces, such a perspective overlooks the increasingly understood ways that targets' group memberships change impressions. When individuals encounter faces from particular social groups, learned stereotypes are automatically activated, even if a perceiver does not endorse them [35].

Gender stereotypes lead women to be viewed more favorably when exhibiting submissive rather than dominant traits. Consequently, trustworthiness becomes more negatively linked to dominance in female faces compared with male faces [26]. As the trustworthiness and dominance dimensions become more correlated for female faces, they collapse into a more homogeneous single dimension among perceivers who strongly endorse gender stereotypes [25]. Stronger anti-Black/pro-White bias is associated with changes in the featural basis of facial trustworthiness, with White-related features conveying trustworthiness and Black-related features conveying untrustworthiness [36]. Older adults are less subject to the negative implications of dominance due to being stereotyped as frail, with dominant older-adult faces conveying wisdom rather than hostility [37].

Increasingly, researchers have used more racially and gender diverse face stimuli to model the structure of face-judgment space directly. Rather than a fixed structure, recent studies have found distinct face-judgment spaces when judging Asian, Black, and White male and female faces, with each race × gender face-judgment space related to distinct racial and gender stereotypes for these groups. Specifically, the extent to which any given pair of traits are deemed more stereotypically related for a given race × gender group predicted a corresponding tendency to judge faces belonging to that group more similarly along the two trait dimensions [38]. For instance, if a perceiver associates being aggressive and being unhealthy more strongly for Black women than for White or Asian women, they are more likely to judge Black women with aggressive-looking facial features as unhealthy. Figure 3 shows network diagrams depicting the structure of

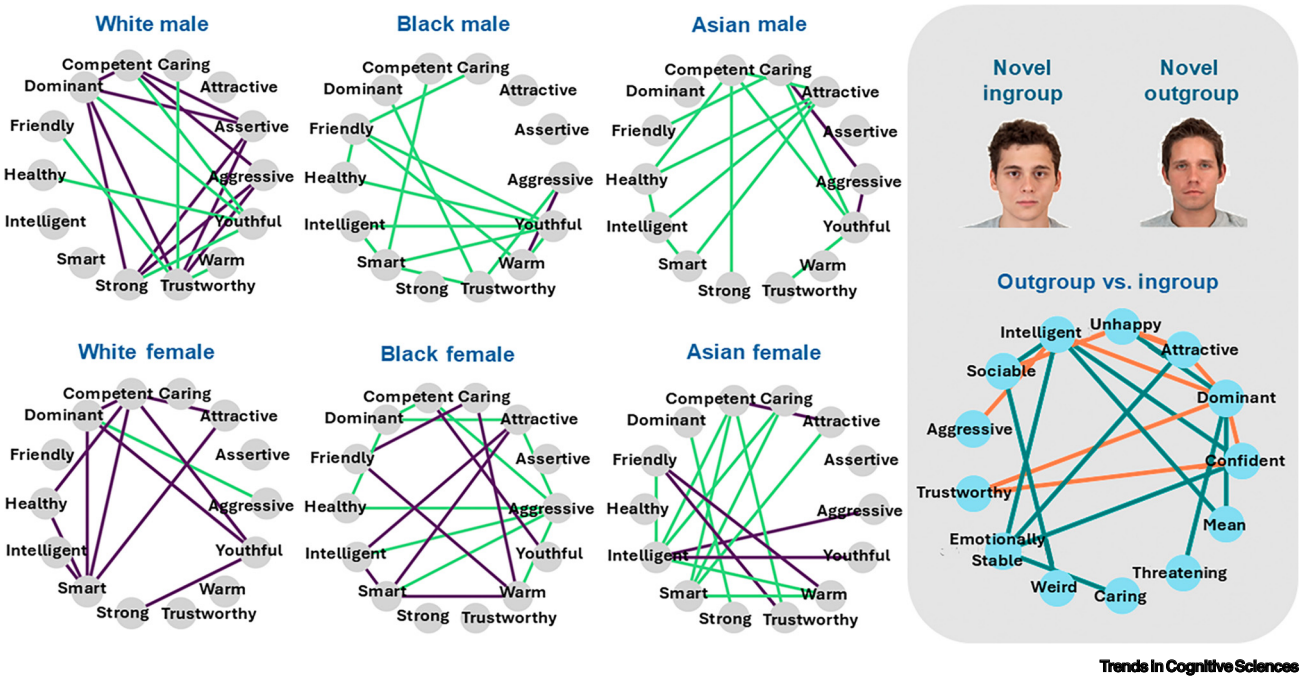

Figure 3. Shifts in face impressions structure across group boundaries. Left: Network diagrams are shown for six race × gender groups, showing how the correlations between facial judgments of every pair of traits became more positive (purple lines) or negative (green lines) depending on the race and gender of the face being judged (relative to the average tendency across the six groups) (data from [38]). Right: A network diagram representing the rapid shift in trait space structure that occurred after assigning perceivers to one of two novel arbitrary groups and asking for impressions of White male faces labeled as part of the perceiver's ingroup or outgroup. Orange lines indicate that the pair of traits became more positively (negatively) correlated when judging ingroup (outgroup) faces; teal lines indicate that the pair of traits became more negatively (positively) correlated when judging ingroup (outgroup) faces (data from [39]). In all diagrams, connections are visualized if the strength (absolute value) of the relative correlation difference |r| > 1 SD for the group. Face images from the Chicago Face Database [54].

face impressions for the six race × gender groups, showing myriad shifts when judging racially and gender-diverse faces. For example, dominance becomes more negatively associated with trustworthiness when judging Black male faces than White male faces [38]. While dimensionality-reduction techniques were able to isolate a relatively similar set of two to three dimensions across the six race × gender spaces, such techniques miss what is systematic variability in face-judgment structure predicted by stereotypes when preserved in a high-dimensional manner.

The groups that perceivers themselves belong to also plays a role. Indeed, prior studies have estimated that perceivers' own race, gender, and interindividual variability are stronger contributors to impressions than facial morphology [28]. Group memberships can be experimentally induced to discover how impressions and their structure rapidly shift. In one set of studies, perceivers were assigned to novel, arbitrary groups (e.g., under-estimators and over-estimators) and judged White male faces labeled as being part of these groups. The classic trustworthiness and dominance dimensions shifted depending on whether the target was labeled an ingroup or outgroup member, becoming more negatively correlated for outgroup members' faces and more positively correlated for ingroup members' faces [39]. Group-based motives can lead dominance on outgroup members to signal intergroup threat and therefore decrease trustworthiness, while on ingroup members to signal prosociality and increase trustworthiness. In our high-dimensional model, all these group-based shifts would reflect the topography of the attractor landscape dynamically adapting to multiple sources of bottom-up and top-down social information (Figure 2).

Concluding remarks

As researchers incorporate more diverse stimuli, participants, and methods, the proliferation of low-dimensional models continues to expand. While they are valuable statistical descriptions of trait judgments, often the literature comes to interpret the relevant dimensions as reflecting distinct functional mechanisms or a privileged cognitive status. The use of factor analysis also moves away from PCA's theoretically agnostic characterizations of judgments to the notion of psychologically meaningful latent constructs that cause those judgments [15]. However, growing evidence of the variability and context dependence of impressions' structure underscores the need for a paradigm shift in our approaches to modeling impressions. The high-dimensional model proposed here treats social impressions as dynamic trajectories through a flexible neural landscape able to be influenced by visual features, conceptual associations, and social cognitive processes.

Our framework complements low-dimensional models by emphasizing that social impressions are highly structured, often in ways that can be fairly consistent due to traits' visuo-semantic similarity or perceivers' chronic goals or internal states. However, it departs from these models by arguing that any low-dimensional model is a mere snapshot of a more dynamic high-dimensional space. With new stimuli, perceivers, or contexts, that snapshot may readily change. While these snapshots may follow relatively consistent patterns, we argue that they need not reflect an intrinsic architecture or latent mechanisms.

Our framework makes several novel predictions, such as: (i) the systematic impact by top-down processes; (ii) the central role of not just featural associations but also conceptual associations; (iii) a proximal-to-distal cascade, where featurally driven traits activate earlier and then propagate to more conceptually driven trait activations; and (iv) hierarchical attractor dynamics, where putatively latent dimensions (e.g., trustworthiness) behave as superordinate attractors that are reached before settling into more granular attractors. It is also important to note that, for participants to rate traits in an experiment, top-down attention helps to activate a focal trait in the model. However, our model assumes that perceivers 'in the wild' can make effectively limitless inferences simultaneously, because such inferences are all tethered to a single high-dimensional neural pattern, primarily based on their visuo-semantic similarity. As such, this framework can answer a classic conundrum of how perceivers act as 'cognitive misers' who strive to maximize efficiency [40] yet can compute seemingly limitless social inferences rapidly in response to another's face [2].

Much challenging work lies ahead to test these ideas, including simulations with recurrent neural networks (as in prior models of the DI theory [20,41,42]) and neuroimaging of faces' multidimensional trait representation [43]. Analyses that probe the structure of representational spaces in their naturally existing high-dimensional form, such as representational similarity analysis [44,45] or graph theory [22,23], have already revealed systematic variability by perceivers, targets, and cultures in impressions' structure not apparent when the space is reduced to a limited set of dimensions [15,24,29,38]. Future research should incorporate high-dimensional analyses and neural-network simulations to better understand the architecture and fluidity of impressions (see Outstanding questions). For researchers simply aiming to identify relevant dimensions for their studies, our framework could also be valuably used to explain and predict the specific low-dimensional structures that are likely to emerge given a particular set of stimuli, perceivers, and context.

Face impressions are often our first reactions and set the stage for more comprehensive perception. We would argue that domain-general principles of attractor dynamics and constraint satisfaction would apply to many forms of social perception [17,21]. For instance, this framework

Outstanding questions

What are the temporal dynamics underlying face impressions? Is there evidence for a proximal-to-distal hierarchy of trait activation, with more featurally driven traits activated earlier and more conceptually driven traits activated later? Time-sensitive techniques [e.g., electroencephalography (EEG)] could be used to answer such questions.

Are there different neural-state spaces based on faces' gender, race, and age, with group stereotypes encoded as differences in conceptual connections between trait nodes, or is there one integrated space with stereotypes encoded by a top-down layer that modulates the attractor landscape of trait activation? Low-dimensional models have often proposed masculinity/femininity or youthfulness as core dimensions, but do these reflect actual trait structure or instead group stereotypes' top-down modulation of trait structure?

Do reduced trait dimensions (e.g., trustworthiness, from PCA or factor analysis) operate as larger, superordinate attractors in a hierarchical relationship with more granular traits? Because neural patterns for reduced dimensions are correlated with neural patterns for a large number of other traits (e.g., sociability, responsibility, caringness), they may activate earlier and pull the system into more granular representations.

Outside a specific task, do indirect measures such as semantic priming or neuroimaging reveal extensive activation of multiple traits simultaneously based on their visuo-semantic similarity? A high-dimensional model argues that perceivers engage in combinations of limitless trait inferences simultaneously, with top-down factors like attentional task demands deepening the attractor basins of states that place maximal focus on traits relevant for the task at hand.

Can multistability processes in attractor neural-network models explain complex impressions involving conflictual traits (e.g., kind but sly) or features (e.g., baby-faced but highly masculine features)? When the system's state is at the boundary between multiple basins of attraction, multistability may occur that involves rapid fluctuation in multiple stable states.

could be leveraged to understand classic Big Two structure, such as warmth and competence, as emergent properties. Moreover, by modeling impressions as a trajectory unfolding over time, our approach could naturally be extended to understand sequences of impressions. In real-world social encounters, by the time a given attractor state has been reached, ongoing nonverbal and verbal cues would already start changing the various attractor states to which the system will start gravitating. Our framework could ultimately be used to model the trajectory of complex impressions during naturalistic encounters. Once further developed, the model could also have several real-world applications. It could be used to better predict when and how biases influence hiring, voting, and legal judgments, improve social interactions, or inform AI technology to more accurately reflect the complexity of human social perception.

For now, we hope that this framework can help to shift the field toward viewing social impressions not so much as drawing on a fixed low-dimensional structure, but as emerging in a combinatorial fashion out of the dynamics of a high-dimensional neural-state space.

Acknowledgments

This work was supported in part by NSF BCS-2235130 (J.B.F.).

Declaration of interests

No interests are declared.

References

1. Freeman, J.B. *et al.* (2014) Amygdala responsivity to high-level social information from unseen faces. *J. Neurosci.* 34, 10573–10581
2. Willis, J. and Todorov, A. (2006) First impressions: making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* 17, 592–598
3. Todorov, A. *et al.* (2015) Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol.* 66, 519–545
4. Zebrowitz, L.A. (2017) First impressions from faces. *Curr. Dir. Psychol. Sci.* 26, 237–242
5. Oosterhof, N.N. and Todorov, A. (2008) The functional basis of face evaluation. *Proc. Natl. Acad. Sci. U. S. A.* 105, 11087–11092
6. Fiske, S.T. *et al.* (2007) Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* 11, 77–83
7. Brambilla, M. *et al.* (2021) The primacy of morality in impression development: theory, research, and future directions. *Adv. Exp. Soc. Psychol.* 64, 187–262
8. Koch, A. *et al.* (2016) The ABC of stereotypes about groups: agency/socioeconomic success, conservative–progressive beliefs, and communion. *J. Pers. Soc. Psychol.* 110, 675–709
9. Abele, A.E. and Wojciszke, B. (2007) Agency and communion from the perspective of self versus others. *J. Pers. Soc. Psychol.* 93, 751–763
10. Vernon, R.J.W. *et al.* (2014) Modeling first impressions from highly variable facial images. *Proc. Natl. Acad. Sci. U. S. A.* 111, E3353–E3361
11. Lin, C. *et al.* (2021) Four dimensions characterize attributions from faces using a representative set of English trait words. *Nat. Commun.* 12, 5168
12. Connor, P. *et al.* (2024) Unconstrained descriptions of Facebook profile pictures support high-dimensional models of impression formation. *Personal. Soc. Psychol. Bull.*, Published online July 30, 2024. https://doi.org/10.1177/01461672241266651
13. Yu, Y. *et al.* (2024) Detecting five-pattern personality traits using eye movement features for observing emotional faces. *Front. Psychol.* 15, 1397340
14. Saucier, G. and Srivastava, S. (2015) What makes a good structural model of personality? Evaluating the big five and alternatives. In *APA handbook of personality and social psychology. Volume 4: personality processes and individual differences*, pp. 283–305, American Psychological Association
15. Jones, B.C. *et al.* (2021) To which world regions does the valence–dominance model of social perception apply? *Nat. Hum. Behav.* 5, 159–169
16. Lu, J. and Lin, C. (2024) From latent constructs to networks: modeling high-dimensional social inferences in naturalistic settings. *PsyArXiv*, Published online November 5, 2024. https://doi.org/10.31234/osf.io/695pm_v1
17. Stolier, R.M. *et al.* (2020) Trait knowledge forms a common structure across social cognition. *Nat. Hum. Behav.* 4, 361–371
18. Martin, A.E. and Slepian, M.L. (2021) The primacy of gender: gendered cognition underlies the Big Two dimensions of social cognition. *Perspect. Psychol. Sci.* 16, 1143–1158
19. Pulvermüller, F. *et al.* (2021) Biological constraints on neural network models of cognitive function. *Nat. Rev. Neurosci.* 22, 488–502
20. Freeman, J.B. and Ambady, N. (2011) A dynamic interactive theory of person construal. *Psychol. Rev.* 118, 247–279
21. Freeman, J.B. *et al.* (2020) Dynamic interactive theory as a domain-general account of social perception. *Adv. Exp. Soc. Psychol.* 61, 237–287
22. Elder, J.J. *et al.* (2023) A fluid self-concept: how the brain maintains coherence and positivity across an interconnected self-concept while incorporating social feedback. *J. Neurosci.* 43, 4110–4128
23. Elder, J. *et al.* (2023) Mapping the self: a network approach for understanding psychological and neural representations of self-concept structure. *J. Pers. Soc. Psychol.* 124, 237–263
24. Stolier, R.M. *et al.* (2018) The conceptual structure of face impressions. *Proc. Natl. Acad. Sci. U. S. A.* 115, 9210–9215
25. Oh, D. *et al.* (2020) Gender biases in impressions from faces: empirical studies and computational models. *J. Exp. Psychol. Gen.* 149, 323–342
26. Sutherland, C.A.M. *et al.* (2015) Face gender and stereotypicality influence facial trait evaluation: counter-stereotypical female faces are negatively evaluated. *Br. J. Psychol.* 106, 186–208
27. Hehman, E. *et al.* (2017) The unique contributions of perceiver and target characteristics in person perception. *J. Pers. Soc. Psychol.* 113, 513–529
28. Xie, S.Y. *et al.* (2019) Perceiver and target characteristics contribute to impression formation differently across race and gender. *J. Pers. Soc. Psychol.* 117, 364–385

29. Oh, D. *et al.* (2022) Personality across world regions predicts variability in the structure of face impressions. *Psychol. Sci.* 33, 1240–1256
30. Bin Meshar, M. *et al.* (2021) Facial stereotyping drives judgments of perceptually ambiguous social groups. *Soc. Psychol. Personal. Sci.* 13, 1221–1229
31. Sutherland, C.A.M. and Young, A.W. (2022) Understanding trait impressions from faces. *Br. J. Psychol.* 113, 1056–1078
32. Sofer, C. *et al.* (2017) For your local eyes only: culture-specific face typicality influences perceptions of trustworthiness. *Perception* 46, 914–928
33. Cuddy, A.J.C. *et al.* (2009) Stereotype content model across cultures: towards universal similarities and some differences. *Br. J. Soc. Psychol.* 48, 1–33
34. Kraft-Todd, G.T. *et al.* (2017) Empathic nonverbal behavior increases ratings of both warmth and competence in a medical context. *PLoS One* 12, e0177758
35. Freeman, J.B. and Johnson, K.L. (2016) More than meets the eye: split-second social perception. *Trends Cogn. Sci.* 20, 362–374
36. Hutchings, R.J. *et al.* (2024) Racial prejudice affects representations of facial trustworthiness. *Psychol. Sci.* 35, 263–276
37. Hehman, E. *et al.* (2014) The face–time continuum: lifespan changes in facial width-to-height ratio impact aging-associated perceptions. *Personal. Soc. Psychol. Bull.* 40, 1624–1636
38. Xie, S.Y. *et al.* (2021) Facial impressions are predicted by the structure of group stereotypes. *Psychol. Sci.* 32, 1979–1993
39. Hong, Y. and Freeman, J.B. (2024) Shifts in facial impression structures across group boundaries. *Soc. Psychol. Personal. Sci.* 15, 619–629
40. Fiske, S.T. and Taylor, S.E. (2020) Social cognition evolves: illustrations from our work on intergroup bias and on healthy adaptation. *Psicothema* 32, 291–297
41. Freeman, J.B. *et al.* (2016) A perceptual pathway to bias: interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychol. Sci.* 27, 502–517
42. Freeman, J.B. *et al.* (2011) Looking the part: social status cues shape race perception. *PLoS One* 6, e25107
43. Chwe, J.A.H. *et al.* (2024) A multidimensional neural representation of face impressions. *J. Neurosci.* 44, e0542242024
44. Freeman, J.B. *et al.* (2018) The neural representational geometry of social perception. *Curr. Opin. Psychol.* 24, 83–91
45. Kriegeskorte, N. *et al.* (2008) Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 4
46. Collova, J.R. *et al.* (2019) Testing the functional basis of first impressions: dimensions for children's faces are not the same as for adults' faces. *J. Pers. Soc. Psychol.* 117, 900–924
47. Twele, A.C. and Mondloch, C.J. (2022) The dimensions underlying first impressions of older adult faces are similar, but not identical, for young and older adult perceivers. *Br. J. Psychol.* 113, 1009–1032
48. Lin, C. *et al.* (2022) How trait impressions of faces shape subsequent mental state inferences. *Hum. Nat. Behav.* 10, m9
49. Lin, C. and Thornton, M. (2023) Evidence for bidirectional causation between trait and mental state inferences. *J. Exp. Soc. Psychol.* 108, 104495
50. Wang, H. *et al.* (2019) A data-driven study of Chinese participants' social judgments of Chinese faces. *PLoS One* 14, e0210315
51. Jones, A.L. and Kramer, R.S.S. (2021) Facial first impressions form two clusters representing approach–avoidance. *Cogn. Psychol.* 126, 101387
52. Sutherland, C.A.M. *et al.* (2013) Social inferences from faces: ambient images generate a three-dimensional model. *Cognition* 127, 105–118
53. DeBruine, L. and Jones, B. (2017) Face Research Lab London Set. *Figshare*, Published online May 28, 2017. https://doi.org/10.6084/m9.figshare.5047666.v1
54. Ma, D.S. *et al.* (2015) The Chicago Face Database: a free stimulus set of faces and norming data. *Behav. Res. Methods* 47, 1122–1135